This paper addresses the problem of optimizing the playback delay experienced by a population of heterogeneous clients, in video streaming applications. We consider a typical broadcast scenario, where clients subscribe to different portions of a scalable video stream, depending on their capabilities. Clients share a common channel, whose limited rate directly drives the playback delays imposed to the different groups of receivers. We derive an optimization problem, that targets a fair distribution of the playback delays among heterogeneous clients. A server–based scheduling strategy is then proposed, that takes into account the properties of the targeted clients, the channel status, and the structure of the media encoding. It is shown to offer significantly reduced playback delays per client population, as compared to traditional scheduling strategies. In the same time, PSNR performance is not affected, which altogether leads to an overall improvement of the quality of service.